# Validation of a common data model for active safety surveillance research

J Marc Overhage,[1,2] Patrick B Ryan,[2,3,4] Christian G Reich,[2] Abraham G Hartzema,[2,5] Paul E Stang[2,3]

[1]Regenstrief Institute, Indiana University, School of Medicine, Indianapolis, Indiana, USA
[2]Observational Medical Outcomes Partnership, Foundation for the National Institutes of Health, Bethesda, Maryland, USA
[3]Johnson & Johnson Institute, Department of Pharmaceutical Research and Development LLC, Titusville, New Jersey, USA
[4]UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, North Carolina, USA
[5]College of Pharmacy, University of Florida, Gainesville, Florida, USA

**Correspondence to**
Dr J Marc Overhage, Regenstrief Institute and Indiana University School of Medicine, Medical Informatics—Regenstrief Institute, Health Information and Translational Sciences Bldg, 410 West 10th Street, Suite 200, Indianapolis, IN 46202, USA; moverhage@regenstrief.org

At the time of this work, AGH was on sabbatical with the Immediate Office of the Commissioner at the US Food and Drug Administration.

Received 16 May 2011
Accepted 16 September 2011
Published Online First 28 October 2011

## ABSTRACT

**Objective** Systematic analysis of observational medical databases for active safety surveillance is hindered by the variation in data models and coding systems. Data analysts often find robust clinical data models difficult to understand and ill suited to support their analytic approaches. Further, some models do not facilitate the computations required for systematic analysis across many interventions and outcomes for large datasets. Translating the data from these idiosyncratic data models to a common data model (CDM) could facilitate both the analysts' understanding and the suitability for large-scale systematic analysis. In addition to facilitating analysis, a suitable CDM has to faithfully represent the source observational database. Before beginning to use the Observational Medical Outcomes Partnership (OMOP) CDM and a related dictionary of standardized terminologies for a study of large-scale systematic active safety surveillance, the authors validated the model's suitability for this use by example.

**Validation by example** To validate the OMOP CDM, the model was instantiated into a relational database, data from 10 different observational healthcare databases were loaded into separate instances, a comprehensive array of analytic methods that operate on the data model was created, and these methods were executed against the databases to measure performance.

**Conclusion** There was acceptable representation of the data from 10 observational databases in the OMOP CDM using the standardized terminologies selected, and a range of analytic methods was developed and executed with sufficient performance to be useful for active safety surveillance.

## BACKGROUND AND SIGNIFICANCE

Observational (non-experimental) studies have several potential advantages over experimental studies (eg, randomized controlled trials), including lower cost, better generalizability, and greater timeliness[1 2]; however, they also have important limitations, including the potential for bias.[3 4] Studies that use observational databases are already a mainstay of drug safety and health-services research, and are viewed as a key resource for comparative effectiveness research. The increasing availability of large, observational, structured healthcare data sets is rapidly increasing the potential for well-designed, observational studies, which can provide valuable insights.

These healthcare data sets are stored in databases that are built using a wide variety of data models and, often, local terminologies. Each of these data models organizes data in a different way, often making it difficult to characterize or analyze the data from disparate healthcare systems in the same way or using the same tools. Not surprisingly, most analyses have focused on data from a single database using a single analysis method customized to the underlying data model and local terminologies. An analysis across multiple disparate databases must either tailor the analysis to accommodate each of the underlying data models and terminologies or convert the databases to a common data model (CDM).

Converting multiple disparate databases to a CDM would allow researchers to write and test the analyses once and then run them on all of the databases with minimal modification. The initial mapping of local codes from each database to standard concepts in a CDM requires detailed knowledge of the local data, but once local codes are converted to the common representation, the requirement for detailed knowledge is minimal. A potential limitation of mapping these individual databases to a CDM is that the CDM may not allow some of the relationships or data contained in a local database to be fully represented. As long as all of the relevant relationships and data required for anticipated uses of the data (drug safety surveillance in our case) are represented, this limitation is not too severe. Terminologies used in various databases vary widely, and selecting a common terminology or set of terminologies is a critical step for implementing a CDM.

Although CDMs have some limitations, conducting studies in disparate databases not converted to a CDM presents substantial challenges. For example, local expertise for each database would be required for each analysis, and the chance for error is increased. In the absence of a common terminology, analyses would have to be customized for each database and would require specific local analytic and database expertise as well as a means to arrive at a rational summary measure across databases.

A CDM, combined with a method for standardizing the terminologies, ensures that analytic methods can be systematically applied to produce meaningfully comparable results, recognizing that differences in the underlying data may produce different results. The prospects for this approach are extremely promising, not only for active safety surveillance but also for other uses such as comparative effectiveness research.

The Observational Medical Outcomes Partnership (OMOP) is empirically assessing the feasibility and utility of using observational data to identify and evaluate associations between medications and

health-related conditions.[5] To facilitate this methodological research, OMOP is evaluating the performance of numerous analytic methods for identifying drug–outcome associations across multiple disparate observational data sources using a CDM[6] and associated standardized terminologies.[7] The OMOP CDM was designed to accommodate data from the observational medical databases that are generally considered necessary for active safety analysis, defining both the format (data model) and the content (standardized terminologies). The OMOP CDM design is intended to serve the purposes of active drug-safety surveillance and to be intuitive, not overly complex, and otherwise 'analyst-friendly.' Before beginning to use the OMOP CDM and a related dictionary of standardized terminologies for a study of large-scale systematic active safety surveillance, we validated the model's suitability for this use by example. To do this, we converted 10 different observational datasets into the OMOP CDM, implemented a number of analytic methods against the OMOP CDM, and executed the methods across all of the databases.

## METHODS

### Model formulation

A set of guiding principles was established to draft a CDM based on expert opinion and previous analyses[8–10] of existing data models.[11–13] The OMOP CDM is a person-centric relational model, with domains inclusive of demographics, observation periods, drug exposure, condition occurrence, procedures, visits, and clinical observations. Reisinger *et al* described the development and initial evaluation of the OMOP CDM.[14]

### Validation through examples

To validate the OMOP CDM's usefulness for active drug-safety surveillance, we examined its ability to represent real-world observational data sets, to provide a convenient conceptual model that facilitates analytic methods development, and to allow the analytic methods to execute quickly enough to be practical. Ten different observational databases (table 1) were used to validate the ability of the OMOP CDM and dictionary of standardized terminologies to represent the diverse data likely to be encountered in drug safety analyses across similar networks of databases. These databases were selected for the validation exercise based on their general appropriateness for drug safety surveillance in terms of the types of data they contain, variability in size, geographic scope, primary purpose for which the data were collected (eg, payment, clinical care, or research), and underlying data models. By converting each of these observational medical databases to the OMOP CDM, we assessed the feasibility, fidelity, and resources required, then subjectively assessed the appropriateness of the CDM to support methods development and executed the methods across all of the observational databases to determine execution time. We validated the suitability of the standardized terminologies to represent the data in the databases (1) by having the medical coders and informaticists performing the mappings identify any fidelity issues they encountered (eg, concepts that could not be represented, concepts that were narrower or broader than the original concepts); (2) by having the analysts identify any fidelity issues they encountered while implementing the analytic methods across a range of medication–outcome pairs; and (3) by examining the effect of alternative standardized terminologies (using the Medical Dictionary for Regulatory Activities (MedDRA), International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM), or Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT) for conditions and National Drug File-Reference Terminology (NDF-RT) or RxNorm for medications) on the results of the analytic methods.

OMOP staff performed the conversion of the five commercially available datasets (GE and the four MarketScan research databases) into the OMOP CDM. For data sources within the distributed network, local teams performed the conversions utilizing their detailed knowledge of the local data structures and coding. All necessary precautions and security measures were implemented to ensure that personal health information was not accessed by non-authorized personnel, including OMOP staff and collaborators. To convert local codes to the standardized terminologies selected for the OMOP CDM, mapping tables from a variety of sources were utilized including publicly available mapping tables (eg, Unified Medical Language System (UMLS) between ICD-9 and SNOMED); mapping tables that are part of the terminology distribution (eg, various drug terminologies and RxNorm); and commercially available mapping tables (eg, Generic Product Identifier and RxNorm). Conversion of local terminologies in the source systems to standardized terminologies represented a challenge. In particular, observations and clinical findings in electronic health records were sometimes identified with local 'codes' that were little more than free text; therefore, string matching of descriptions (eg, Veteran's Administration Product and RxNorm) and manual coding (performed and validated by Health Language) were employed.

OMOP created two software programs, the Observational Source Characteristics Analysis Report (OSCAR)[15] and the Generalized Review of OSCAR Unified Checking (GROUCH),[16] to characterize the data once they were converted to the OMOP CDM and to identify potential issues across all OMOP CDM tables, including potential concerns with all drug exposures and conditions. These programs allow the data to be compared across databases and with local knowledge about the data sources to validate the conversion process.

OSCAR allows the creation of summary statistics about the data that could be compared with similar summary statistics generated from the source data and other data sets. These summary statistics include the number of persons, number of conditions, number of medications, and duration of data capture, each broken down by age and gender; proportions of medications and conditions per patient; and others.

GROUCH identifies potential data anomalies across all OMOP CDM tables, including potential concerns with all drug exposures and all conditions, by testing for statistically significant variations across a large number of relationships and for clinically unexpected findings such as prostate cancer in females and pregnancy in males. This allows for data-quality review of specific medications (such as the OMOP medications of interest) or specific conditions (including population-level prevalence of the health outcomes of interest and gender-stratified rates).

GROUCH organizes potential data anomalies into three categories: concept, boundary, and temporal warnings. Concept issues are related to the observed proportion of patients with which a concept appears in a data source compared with the observed proportions in other data sources. Boundary issues include the appearance of unusual (suspicious or implausible) values such as a year of birth greater than the current year, an age >110 years, a number of days of medication supply less than zero, or a length of drug era (period of time a patient is inferred to be continuously exposed to

**Table 1** Summary of the 10 observational databases used to validate the Observational Medical Outcomes Partnership common data model and dictionary of standardized terminologies, including a brief description, the approximate population size, and the terminologies used

| Name | General database description | Population size (M) used for validation* | Terminologies | |
|---|---|---|---|---|
| | | | Claims | Clinical |
| GE centricity EMR database | Derived from data pooled by providers using GE Centricity Office (an ambulatory electronic health record) into a data warehouse in a HIPAA-compliant manner | 11.2 | | ICD-9<br>CPT-4<br>local |
| MarketScan research databases from Thomson Reuters | MarketScan Lab Database—represents privately insured population, with administrative claims from inpatient, outpatient, and pharmacy services supplemented by laboratory results | 1.5 | ICD-9<br>CPT-4<br>HCPCS<br>NDC | LOINC |
| | MarketScan Medicaid Multi-State Database—contains administrative claims data for Medicaid enrollees from multiple states | 11.1 | ICD-9<br>CPT-4<br>HCPCS<br>NDC | |
| | MarketScan Medicare Supplemental and Coordination of Benefits Database—captures administrative claims for retirees with Medicare supplemental insurance paid by employers, including services provided under Medicare-covered payment, employer-paid portion, and any out-of-pocket expenses | 4.4 | ICD-9<br>CPT-4<br>HCPCS<br>NDC | |
| | MarketScan Commercial Claims and Encounters—represents privately insured population and captures administrative claims with patient-level de-identified data from inpatient and outpatient visits and pharmacy claims of multiple insurance plans | 58 | ICD-9<br>CPT-4<br>HCPCS<br>NDC | |
| Humana | Contains medical (inpatient, outpatient, and emergency room), pharmacy, and laboratory data (including test results) from Humana's administrative claims database of medical members | 6.5 | ICD-9<br>CPT-4<br>HCPCS<br>NDC | Local |
| i3 Drug safety, ingenix normative health information | Contains healthcare claims from across the USA with medical and pharmacy benefit coverage where available | 1 | ICD-9<br>CPT-4<br>HCPCS<br>NDC | LOINC |
| Partners healthcare system | Includes data from the partner's clinical transaction-based data repository as well as inpatient and outpatient billing feeds being collected in the research patient data registry, an analytic-structured database | 3 | | ICD-9<br>CPT-4<br>local |
| Regenstrief Institute/ Indiana Network for patient care | Includes healthcare data from the Indiana Network for Patient Care containing population-based, longitudinal, and structured coded and text data captured from hospitals, physician practices, public health departments, laboratories, radiology centers, pharmacies, pharmacy benefit managers, and payers | 2 | ICD-9<br>CPT-4<br>HCPCS<br>NDC | ICD-9<br>CPT-4<br>LOINC<br>SNOMED<br>RxNorm<br>NDC<br>local |
| SDI health | Contains HIPAA-compliant, deidentified, encrypted patient-level data from hospitals, clinics, physician offices, and retail and specialty pharmacies from all 50 US states | 90 | ICD-9<br>CPT-4<br>HCPCS<br>NDC | |

*For some data sources, data from a subset of patients available in the source were incorporated into the Observational Medical Outcomes Partnership common data model instance.
CPT-4, Current Procedural Terminology, 4th edn; EMR, electronic medical record; HCPCS, Healthcare Common Procedure Coding System; HIPAA, Health Insurance Portability and Accountability Act; ICD-9, International Classification of Diseases, Ninth Revision; LOINC, Logical Observation Identifiers Names and Codes; NDC, National Drug Code; SNOMED, Systematized Nomenclature of Medicine.

a medical product) longer than the period of time for which the patient was observed. Temporal issues include anomalies in patterns such as unusual spikes in the monthly rate of drug records per person over time or a significant change at a particular year of birth within the age distribution. Together, these categories cover the types of anomalies (incomplete, implausible, and suspicious) at all levels and across all tables within the OMOP CDM. Any anomaly-detection approach identifies potential concerns, but not all issues identified represent legitimate anomalies. Instead, many of the potential issues will be determined to be true artifacts of the data that require no action beyond additional specification to clearly acknowledge and communicate the justification for why the observed anomaly is acceptable.

OMOP staff, methods collaborators, and participants in the OMOP Cup (a methods competition)[17] developed analytic methods to be executed across all databases to assess whether the OMOP CDM sufficiently supports efficient and feasible analysis. The OMOP staff created analyses based on methods that have been described for active safety surveillance and epidemiological studies; methods collaborators developed novel methods or extensions to established methods; and OMOP Cup participants developed innovative and exploratory methods. These methods represent a broad array of known approaches that might be used for active safety surveillance, including dozens of variations. The methods took advantage of the semantic network that is part of the dictionary to aggregate both medications and outcomes.

## RESULTS

The multiple disparate data sources were all successfully converted to the OMOP CDM, including all observational data elements that are relevant to identifying drug exposures and defining condition occurrences. In part, this was because the observations table allows almost any data to be represented in a generic fashion using the Entity-Attribute-Value (EAV) model which is found frequently in large clinical data repositories.[18]

In fewer than 10 cases, the experts creating the mapping felt that the concepts available in the standardized terminologies that are part of the OMOP CDM were more specific than the concept from the source terminology that they were trying to map. This mismatch in specificity raised the question of how best to map these data elements, particularly since mapping to a 'higher level' or more general code in the standardized terminologies may result in 'lumping' the result with other results. These differences in specificity are a predictable result of variation in the granularity of the source data.

Not all available patients were included in the OMOP CDM instantiation created by each distributed research partner. Several factors contributed to the choice of patients to include, including time and computing-resource constraints, exclusion of certain data from military and other large clients, exclusion of dental and supplemental insurance claims, and the completeness of longitudinal data available.

As noted earlier, the five data sources within the distributed network were transformed by the corresponding distributed research partner. On average, converting a database to the OMOP CDM, including mapping terminologies, required the equivalent of four full-time employees for 6 months and significant computational resources for each distributed research partner. Each partner utilized a number of people with a wide range of expertise and skills to complete the project, including project managers, medical informaticists, epidemiologists, database administrators, database developers, system analysts/programmers, research assistants, statisticians, and hardware technicians. Knowledge of clinical medicine was critical to correctly map data to the proper OMOP CDM tables. The five commercial databases were transformed by the same OMOP team, which required knowledge of a broader set of data but also gained efficiencies in establishing shared processes that could be applied across sources. In addition, the commercial databases were more fully normalized.

Computing resource requirements for converting the databases varied widely, in part because of the amount of data available to each distributed research partner. The size and complexity of the data transformation meant that the procedures had to run for long periods of time. Load and run times for conversion ranged from 4 to 11 days, and conversions typically ran on a quad-core server.

After OMOP CDM conversion, the characterizations of the databases produced by OSCAR and GROUCH compared well with expectations, suggesting that the conversion process performed as expected. For the five commercial databases, detailed examination of summary statistics derived from OSCAR and from the source observational databases initially identified a few issues, most of which were corrected in subsequent conversions. These issues included zip codes 00001 through 00009 being handled incorrectly; a few procedure codes that were not correctly mapped to medications; a rounding error for drug quantities less than one; the drug exposure length being incorrectly programmed, which resulted in erroneous values in 3.72% of cases; and the condition era length being incorrectly programmed, which resulted in a small number of erroneous values. The distributed partners also reviewed the database characterizations and identified a number of issues, some of which were attributable to anomalies in the original data sources but only a few of which represented errors in data handling. Most represented explainable phenomenon based on local practices or idiosyncrasies.

We corrected errors in the conversion process in the few cases where they were identified. For data anomalies identified by GROUCH that were not determined to be the result of programming errors, we did not apply additional rules for data manipulation. Instead, we preserve source data in the CDM (even if they may contain potentially erroneous values), and

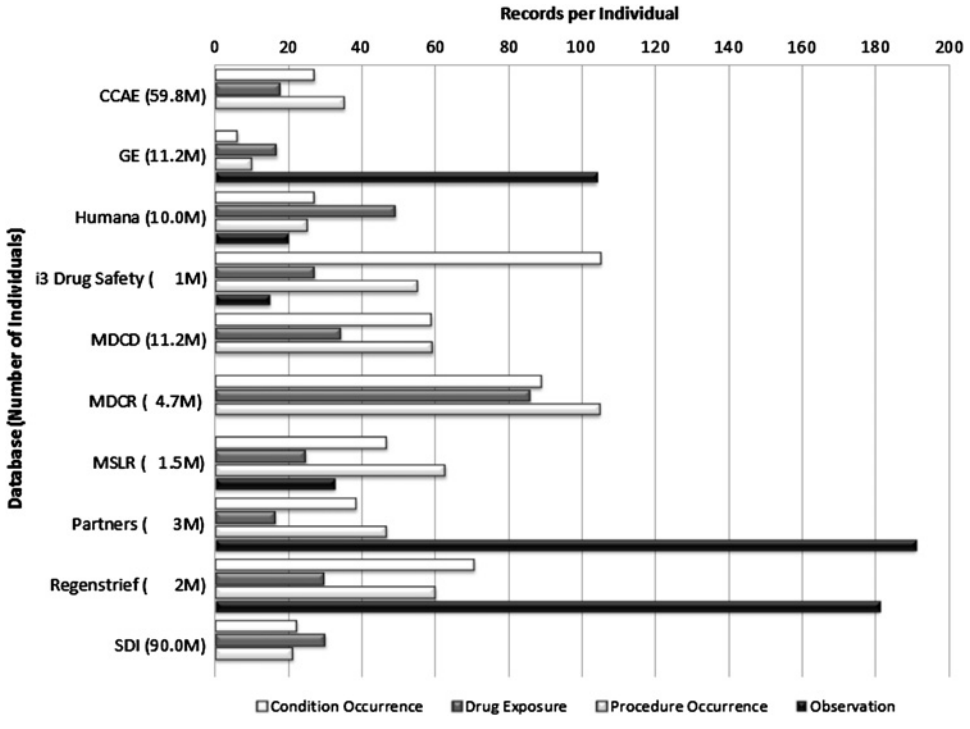


**Figure 1** Database characteristics. Rates of conditions, medications, procedures, and observations per person varied widely across databases. CCAE, Commercial Claims and Encounters; MDCD, MarketScan Medicaid Multi-State Database; MDCR, Medicare Supplemental and Coordination of Benefits Database; MSLR, MarketScan Lab Database.

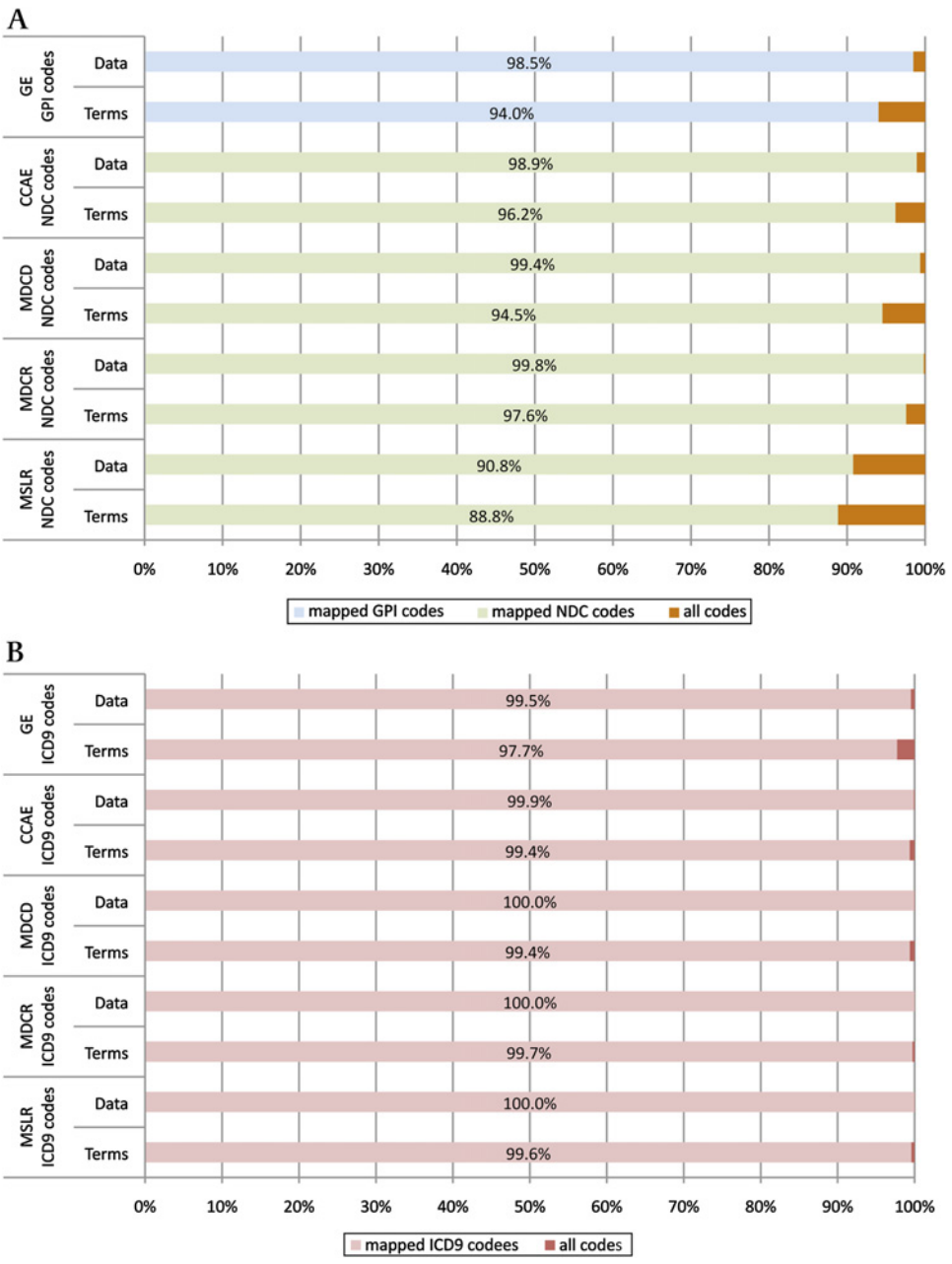


**Figure 2** Proportion of terms and database records for drugs (A) and medications (B) that could be mapped using different standard terminologies for the five commercial databases, demonstrating the suitability of the standardized terminologies chosen for the OMOP CDM. CCAE, Commercial Claims and Encounters; GPI, generic product identifier; ICD9, International Classification of Diseases, Ninth Revision; MDCD, MarketScan Medicaid Multi-State Database; MDCR, Medicare Supplemental and Coordination of Benefits Database; MSLR, MarketScan Lab Database; NDC, National Drug Code.

defer additional data cleaning to the analysis phase, since we believe more informed and context-specific decisions can be made at that point.

To examine the suitability of the standardized terminologies chosen for the OMOP CDM, we characterized the number of concepts and the number of database records that the system could map. Figure 1 summarizes the approximate size of the data tables created from each database. For the five commercial databases, the percentage of concepts mapped ranged from 91.8% to 99.6% for conditions and 56.1% to 74.8% for medications, and the percentage of database records mapped ranged from 93.2% to 99.7% for conditions and 88.8% to 97.6% for medications (figure 2). Figure 3 shows the frequency, as a percentage of total records, of concepts that appear in a database at a rate more than three standard deviations from the mean rate across all databases, ie, are a lot more frequent in comparison to the community of databases. Only two concepts were used in more than 0.10% of records.

Eleven different statistical methods with dozens of parameterized variations were created and executed against 10 different OMOP CDM instantiations for several hundred different drug–outcome pairs. In addition, OMOP Cup participants created methods that analyzed the data in OMOP CDM format. In total, more than 75 individuals were involved in creating the methods, and although they asked clarifying questions about the data model, none found the model confusing or limiting in their work. In fact, subjectively they found that the model directly facilitated development of the methods. All methods developers, for example, took advantage of drug and condition eras rather than deriving their own. None of the developers encountered any limitations as a result of the standardized terminologies chosen for the OMOP CDM. The methods were executed without modification (although the configuration of the execution environment and availability of specific software packages limited which methods a distributed research partner could execute). Table 2 provides the means and standard

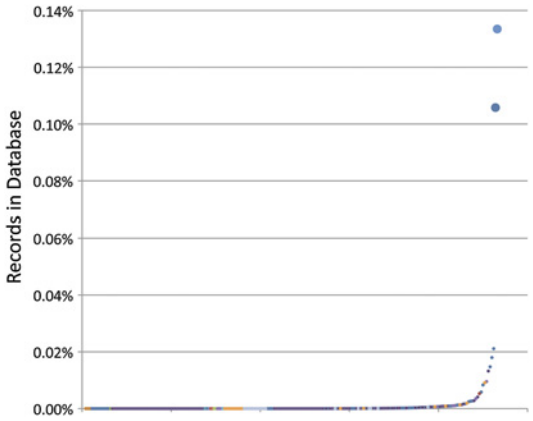


**Figure 3** Graph showing the frequency, as a percentage of records, with which concepts that appear in a database at a rate more than three deviations from the mean frequency computed across all databases. Only two concepts (the RxNorm code for amlodipine 10 mg/benazepril 20 mg oral capsule and the Systematized Nomenclature of Medicine code for large liver) appeared in more than 0.10% of the records in a database. CCAE, Commercial Claims and Encounters; GPI, generic product identifier; ICD9, International Classification of Diseases, Ninth Revision; MDCD, MarketScan Medicaid Multi-State Database; MDCR, Medicare Supplemental and Coordination of Benefits Database; MSLR, MarketScan Lab Database; NDC, National Drug Code.

deviations of execution times of the 11 statistical methods across all databases (excluding i3 Drug Safety). The maximum mean execution time was 34.39±37.16 h, and the minimum was 0.81±0.91 h. Executing the methods using alternative standardized terminologies to aggregate concepts or medications did not significantly affect the point estimates (data not shown).

## DISCUSSION

We were able to validate the OMOP CDM and standardized terminologies, demonstrating that they accommodated a broad range of observational data (both administrative claims and electronic health records) and supported the development of analytic methods and that these analyses executed in a manner efficient enough to be useful for active drug safety surveillance and other, similar analyses.

**Table 2** Execution time of the 11 standardized methods across nine databases (results were not available from Ingenix Normative Health Information database)

| Method name* | No of parameter combinations | Execution time (mean±SD in hours) |
|---|---|---|
| Observational screening | 162 | 0.81±0.91 |
| Univariate self-controlled case series | 64 | 1.83±2.24 |
| High-throughput Safety screening Indiana University population-based method | 6 | 2.72±3.07 |
| Multi-set case control estimation | 32 | 2.73±2.5 |
| Information Component-based measure of disproportionality temporal pattern | 84 | 3.16±5.28 |
| Bayesian logistic regression | 24 | 6.64±8.77 |
| Disproportionality analysis | 112 | 9.63±8.88 |
| High-dimensional propensity score | 144 | 10.16±12.89 |
| Case-crossover | 48 | 11.41±12.72 |
| Maximized sequential probability ratio test | 144 | 21.3±16.67 |
| Conditional sequential sampling procedure | 144 | 34.39±37.16 |

Time represents one run of the method, averaged across all parameter combinations and across 235 drug–outcome pairs.
*More details available at http://omop.fnih.org/Methods library (accessed 1 Aug 2011).

Several strengths were identified in the CDM approach and the OMOP CDM in particular. The creation of a CDM allowed geographically dispersed, collaborating researchers to understand the data and minimized confusion about how the data were organized. Similar analyses carried out by different organizations in their distinct databases may allow researchers to begin to understand how analysis results depend on the data used. Given that the analysis methods, terminologies, and data model are identical, differences are more likely to be due to the underlying patient populations and the data captured about them which, as illustrated in figure 1, can be quite significant. This improved consistency should improve the power of synthetic methods such as meta-analysis to extract additional insights. Based on the experience of methods developers, inferred observation periods for creation of medication or condition eras seem to be an appropriate approach and simplify both characterizing the data and development of analysis methods. In addition, several potential weaknesses were identified. We purposely chose a 'least common denominator' data model that accommodated all of the data and was relatively easy to translate a variety of data models into, but at the expense of some loss of richness. Aggregation of clinical observations by encounter that might exist in a source database, for example, would be lost. We did not find in any of our validation examples that reducing the complexity of the data limited its use for drug-safety surveillance.

Advantageously, the model provides clear and explicit specification for a standard terminology for each database table in the OMOP CDM, including SNOMED for condition occurrences, RxNorm for medications, and Logical Observation Identifiers Names and Codes (LOINC) for results/clinical observations. Despite this benefit, we remained concerned about the impact of the evolution of terminologies over time and recommend that careful attention be devoted to version management to avoid the situation in which a source code could be mapped incorrectly if it was an older code, and 'current' mappings were used.

Another important problem was encountered when dealing with lab values. Different units may be used for the same laboratory test, not all of which are reflected in the Unified Codes for Units of Measure (UCUM) terminology; similarly, some results codes for highly specialized laboratory tests are not represented by LOINC. These issues do not occur very often, but they could have an impact on particular analyses because these highly specialized lab tests would not be stored in the OMOP CDM using the chosen terminologies.

The OMOP CDM and dictionary required that related data be represented in a common location in the database. In an observational database, data about medication exposure, for example, might be found in billing codes (eg, Healthcare Common Procedure Coding System) for chemotherapy administration, as drug-dispensing events from claims data, or in medication-order records. When the data were converted to the OMOP CDM, all drug exposures were aggregated in a single table, no matter what data in the source database identified the exposure.

An important consideration with the use of the OMOP CDM is the possible loss of granularity of data by forcing disparate sources into one common model. In particular, when mapping between concepts in different terminologies, various types of complexities may be encountered. First, the source concept may not map to *any* target concept, such as when the source terminology has a broader concept coverage than the target terminology. This would result in more than one target code and ambiguous mappings that are not one-to-one translations. For example, the ICD-9-CM code 284 (aplastic anemia and other bone marrow failure syndromes) was explicitly excluded from

a definition of aplastic anemia created for analysis, but when utilizing mappings to the MedDRA term marrow depression and hypoplastic anemias (HLT), code 284 was included.

Second, a standardized approach using a generic data model may obscure some of the details that could potentially make an analysis richer. Specifically, if terminology is used slightly differently across different databases, then nuances in mapping to standardized terminologies may imply unwarranted assumptions. The OMOP CDM anticipated this issue, allowing the original codes to be stored along with the standardized codes so that subsequent analyses could take these codes into account.

As a final point, researchers often create 'purpose built,' handcrafted databases for each study, after thinking carefully about database content and structure. Analytic criteria applied may be study-specific, and it might not be possible to replicate this specificity using a systematic analysis approach based on a CDM. OMOP CDM structural limitations were expected to possibly inhibit adequate representation of the observational data; however, we did not find any examples in which this happened, though we did not attempt to represent encounter-level organization of the data. This constraint did not inhibit any methods developers, and ensuing methods did not draw on encounter-level organization of the data.

Converting data to the OMOP CDM required significant effort, a broad range of expertise, and extensive computational resources. We underestimated the level of effort and the breadth of skill sets (especially data mapping) required by the OMOP distributed research partners to convert their data to the OMOP CDM. There were several reasons why additional effort and skills were required, including the partners having to understand the OMOP CDM and to conceptualize how to rationalize the OMOP CDM with their data model or models (some partners had data in more than one system). Also, there were a myriad of issues related to mapping local terminologies to the standardized terminologies. In addition, the broad range of expertise and knowledge necessary to successfully convert the data required the participation of individuals from different parts of the organization, increasing the effort required for coordination. We believe that this level of effort will only be required for the initial conversion effort for a data source and that subsequent conversions will require significantly fewer resources, since most of the effort is reusable. While there is some potential to reduce the effort required for mapping terminologies, particularly as standards are more broadly adopted, there is little opportunity to reduce the overall effort, as much of the necessary work required involved externalizing knowledge about local systems. The complexity of this task is demonstrated, in part, through the diversity of skill sets required to perform the conversion to the OMOP CDM.

Some individuals found that representing their data in the OMOP CDM challenged some of the assumptions that they usually made for traditional epidemiologic studies, such as focusing on a previously specified medication exposure—health outcome of interest pair and the need for customized data cleaning programs. Often, these types of tacit knowledge are stored only in the minds of frequent users of the data. For example, the reliability of data in one table or from one source may be very different than the reliability of data from another table or source, but this knowledge is not represented in the data model. An additional benefit of converting data sources to the OMOP CDM and generating detailed descriptive statistics for each of them is that these statistics provided insights into differences in the data facilitated subsequent identification of problems in the data sources or in the conversion process including mapping.

## CONCLUSION

We validated the OMOP CDM and integrated standardized terminologies for active safety surveillance using 10 example databases and 11 analytical methods. Using a CDM creates a basis for broad collaboration among researchers and practitioners, and, in particular, allows methods and database developers to work with a high degree of autonomy based on a shared conceptual model. The combination of data represented in a CDM and tools designed and tested against the CDM offers compelling potential for enhancing the rate of development and application of an active safety-surveillance system.

**Acknowledgments** The authors thank and acknowledge the contributions of the OMOP distributed research partners in phases of this research, who are supported by a grant from the Foundation for the National Institutes of Health.

**Funding** The Observational Medical Outcomes Partnership is funded by the Foundation for the National Institutes of Health through generous contributions from the following: Abbott, Amgen, AstraZeneca, Bayer Healthcare Pharmaceuticals, Bristol-Myers Squibb, Eli Lilly & Company, GlaxoSmithKline, Johnson & Johnson, Lundbeck, Merck & Co., Novartis Pharmaceuticals Corporation, Pfizer, Pharmaceutical Research Manufacturers of America (PhRMA), Roche, sanofi-aventis, Schering-Plough Corporation, and Takeda.

**Competing interests** None.

**Provenance and peer review** Not commissioned; externally peer reviewed.

## REFERENCES

1. **Benson K,** Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med* 2000;**342**:1878—86.
2. **Concato J,** Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med* 2000;**342**: 1887—92.
3. **Moses LE.** Measuring effects without randomized trials? Options, problems, challenges. *Med Care* 1995;**33**(4 suppl):AS8—14.
4. **Kunz R,** Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ* 1998;**317**:1185—90.
5. **Stang PE,** Ryan PB, Racoosin JA, *et al*. Advancing the science for active surveillance: rationale and design for the observational medical outcomes partnership. *Ann Intern Med* 2010;**153**:600—6.
6. **Observational Medical Outcomes Partnership [Internet].** *Common Data Model, ETL Process, & Terminologies; c2009—2011*. 2009. http://omop.fnih.org/ETLProcess (accessed 25 Jul 2011).
7. **Observational Medical Outcomes Partnership [Internet].** *OMOP Standardized Terminologies; c2009—2011*. http://omop.fnih.org/vocabularies (accessed 25 Jul 2011).
8. **Brown J,** Lane K, Moore K, *et al*. *Database Models to Implement the FDA Sentinel Initiative. Final Report [Internet]*. 2009. FDA-2009-N-0192—005. http://www.regulations.gov/#!documentDetail;D=FDA-2009-N-0192-0005 (accessed 5 Mar 2010).
9. **Ryan PB,** Mera R, Merrill GH. *Opportunities and Challenges in Leveraging Observational Data for Pharmacovigilance*. Washington, DC: AMIA Pharmacovigilance and Informatics Summit, 2007.
10. **Gliklich RE,** Dreyer NA, eds. *Registries for Evaluating Patient Outcomes: A User's Guide. Prepared by Outcome DEcIDE Center, AHRQ Publication No 07-EHC001-1*. Rockville (MD): Agency for Healthcare Research and Quality, 2007.
11. **Health Level Seven International [Internet].** *HL7 Reference Information Model; c2007—2011*. http://www.hl7.org/implement/standards/rim.cfm (accessed 18 Nov 2010).
12. *i2b2: Informatics for Integrating Biology & the Bedside [Internet]*. https://www.i2b2.org (accessed 18 Nov 2010).
13. *HMO Research Network [Internet]. c2011*. http://www.hmoresearchnetwork.org (accessed 18 Nov 2010).
14. **Reisinger SJ,** Ryan PB, O'Hara DJ, *et al*. Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. *J Am Med Inform Assoc* 2010;**17**:652—62.
15. **Observational Medical Outcomes Partnership [Internet].** *OSCAR—Observational Source Characteristics Analysis Report (OSCAR) Design Specification and Feasibility Assessment; c2009—2011*. http://omop.fnih.org/OSCAR (accessed 3 Dec 2010).
16. **Observational Medical Outcomes Partnership [Internet].** *Generalized Review of OSCAR Unified Checking; c2009—2011*. http://omop.fnih.org/GROUCH (accessed 11 Jan 2011).
17. **Observational Medical Outcomes Partnership [Internet].** *OMOP Cup; c2009—2011*. http://omop.fnih.org/omopcup (accessed 11 Jan 2011).
18. **Dinu V,** Nadkarni P. Guidelines for the effective use of entity-attribute-value modeling for biomedical databases. *Int J Med Inform* 2007;**76**:769—79. doi:10.1016/j.ijmedinf.2006.09.023.